11-2007

Improving the Reliability of Causal Discovery from Small Data Sets using the Argumentation Framework

Facundo Bromberg
Iowa State University

Dimitris Margaritis
Iowa State University

Improving the Reliability of Causal Discovery from Small Data Sets using the Argumentation Framework

Abstract
We address the problem of reliability of independence-based causal discovery algorithms that results from unreliable statistical independence tests. We model the problem as a knowledge base containing a set of independences that are related through the well-known Pearl's axioms. Statistical tests on finite data sets may result in errors in these tests and inconsistencies in the knowledge base. Our approach uses an instance of the class of defeasible logics called argumentation, augmented with a preference function that is used to reason and possibly correct errors in these tests, thereby resolving the corresponding inconsistencies. This results in a more robust conditional independence test, called argumentative independence test. We evaluated our approach on data sets sampled from randomly generated causal models as well as real-world data sets. Our experiments show a clear advantage of argumentative over purely statistical tests, with improvements in accuracy of up to 17%, measured as the ratio of independence tests correct as evaluated on data. We also conducted experiments to measure the impact of these improvements on the problem of causal structure discovery. Comparisons of the networks output by the PC algorithm using argumentative tests versus using purely statistical ones show significant improvements of up to 15%.

Disciplines
Artificial Intelligence and Robotics

Improving the Reliability of Causal Discovery from Small Data Sets using the Argumentation Framework

Facundo Bromberg
Dept. of Computer Science
Iowa State University
Ames, IA 50011
bromberg@cs.iastate.edu

Dimitris Margaritis
Dept. of Computer Science
Iowa State University
Ames, IA 50011
dmarg@cs.iastate.edu

November 2007

TR–ISU–CS–07–15

Abstract

We address the problem of reliability of independence-based causal discovery algorithms that results from unreliable statistical independence tests. We model the problem as a knowledge base containing a set of independences that are related through the well-known Pearl's axioms. Statistical tests on finite data sets may result in errors in these tests and inconsistencies in the knowledge base. Our approach uses an instance of the class of defeasible logics called argumentation, augmented with a preference function that is used to reason and possibly correct errors in these tests, thereby resolving the corresponding inconsistencies. This results in a more robust conditional independence test, called argumentative independence test. We evaluated our approach on data sets sampled from randomly generated causal models as well as real-world data sets. Our experiments show a clear advantage of argumentative over purely statistical tests, with improvements in accuracy of up to 17%, measured as the ratio of independence tests correct as evaluated on data. We also conducted experiments to measure the impact of these improvements on the problem of causal structure discovery. Comparisons of the networks output by the PC algorithm using argumentative tests versus using purely statistical ones show significant improvements of up to 15%.

1 Introduction and Motivation

Directed graphical models, also called Bayesian networks, are used to represent the probability distribution of a domain. This makes them a useful and important tool for machine learning, where a common task is predicting the probability distribution of a variable of interest given some other knowledge, usually in the form of values of other variables in the domain. An additional use of Bayesian networks comes by augmenting them with additional causal semantics that represent cause and effect relationships in the domain. The resulting networks are called causal. An important problem is inferring the structure of these networks, a process that is sometimes called *causal discovery*, which can provide to a researcher insights into the underlying data generation process.

Two major classes of algorithms exist for learning the structure of Bayesian networks. One class, which contains so-called *score-based* methods, learns the structure by conducting a search in the space of all structures trying to find the maximum of a score function, which is usually a penalized log-likelihood e.g., the Bayesian information criterion or the (equivalent) minimum description length. Through the use of such score functions, algorithms in this class address the problem of causal discovery indirectly, focusing instead on accurate prediction. A second class instead works by exploiting the fact that a causal Bayesian network implies the existence of certain conditional independence statements between subsets of the domain variables. Algorithms in this class use the result of a number of conditional independences to constrain the set of possible structures consistent with these to a singleton (if possible) and infer that structure as the only possible one. As such they are called *constraint-based* or *independence-based* algorithms. In this paper we address open problems related to the latter class of algorithms.

It is well-known that independence-based algorithms have several shortcomings. A major one has to do with the effect that unreliable independence information has on the their output. In general such independence information comes from two sources: (a) a domain expert that can provide his or her opinion on the validity of certain conditional independences among some of the variables, usually with a degree of confidence attached to them, and/or (b) statistical tests of independence, conducted on data gathered from the domain. As expert information is often costly and difficult to obtain, the latter is the most commonly used option in practice. A problem that occurs frequently however is that the data set available may be small. This may happen for various reasons: lack of subjects to observe (e.g., in medical domains), expensive data-gathering process, privacy concerns and others. Unfortunately, the reliability of statistical tests significantly diminishes on small data sets. For example, Cochran (1954) recommends that Pearson's χ^2 test is deemed unreliable if more than 20% of the cells of the test's contingency table have an expected count of less than 5 data points. Unreliable tests, besides producing errors in the resulting causal model structure, may also produce cascading errors due the way that independence-based algorithms work: their operation, including which test to evaluate next, typically depends on the outcomes of previous ones. Therefore, an error in a previous test may have large (negative) consequences in the resulting structure, a property that is called algorithm *instability* in Spirtes et al. (2000). In this paper we present a number of methods for increasing the reliability of independence tests for small data sets and, as a result, the reliability of independence-based algorithms that use them.

We model this setting as a propositional knowledge base whose contents are conditional independences that are potentially inconsistent. Our main insight is to recognize that the outcomes of independence tests are not themselves independent but are constrained by the outcomes of other tests through Pearl's well-known properties of the conditional independence relation (Pearl, 1988). Therefore, such constraints can be sometimes used to correct certain inconsistent test outcomes, choosing instead the outcome that can be inferred by other tests that are not involved in contradictions. We illustrate this by an example.

Example 1. *Consider an independence-based knowledge base that contains the following propositions, obtained through statistical tests on data.*

$$(\{0\} \perp\!\!\!\perp \{1\} \mid \{2,3\}) \tag{1}$$
$$(\{0\} \perp\!\!\!\perp \{4\} \mid \{2,3\}) \tag{2}$$
$$(\{0\} \not\perp\!\!\!\perp \{1,4\} \mid \{2,3\}) \tag{3}$$

where $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})$ *denotes conditional independence of the set of variables* $\mathbf{X}$ *with* $\mathbf{Y}$ *conditional on set* $\mathbf{Z}$*, and* $(\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})$ *denotes conditional dependence. Suppose that (3) is in fact wrong. Such an error can be avoided if there exists some constraint involving these independence propositions. For example, suppose that we also know that the following rule holds in the domain (this is an instance of the Composition axiom, described later in Section 2).*

$$(\{0\} \perp\!\!\!\perp \{1\} \mid \{2,3\}) \wedge (\{0\} \perp\!\!\!\perp \{4\} \mid \{2,3\}) \implies (\{0\} \perp\!\!\!\perp \{1,4\} \mid \{2,3\}). \tag{4}$$

We assume that such rules correspond to theoretical domain properties and are always correct (see next section for more details). Rule (4) and dependence proposition (3) contradicts each other, resulting in an inconsistent knowledge base. Therefore proposition (3) can no longer be accepted. The incorrect independence of proposition (3) could be rejected (and the error corrected) if it was possible to resolve the inconsistency in favor of implication (4). The framework presented in the rest of the paper provides a principled approach for resolving such inconsistencies.

The situation described in the previous example, while simple, demonstrates the general idea that we will use in the rest of the paper: the set of independences and dependences used in a causal discovery algorithm form a potentially inconsistent knowledge base, and making use of general rules that we know hold in the domain helps us correct certain outcomes of statistical tests from (frequently more than one) other ones. In this way we will be able to improve the reliability of causal discovery algorithms that use them to derive causal models. To accomplish this we will use the framework of *argumentation*, which provides a sound and elegant way of resolving inconsistencies in such knowledge bases, including ones that contain independencies.

The rest of the paper is organized as follows. The next section introduces our notation and definitions. Section 3 presents the argumentation framework and its extension with preferences, and describes our approach for applying it to represent and reason in potentially inconsistent independence knowledge bases. We present our experimental evaluation in Section 4, and conclude with a summary of our approach and possible directions of future research in Section 5.

2 Notation and Preliminaries

In this work, we denote random variables with capitals (e.g., X, Y, Z) and sets of variables with bold capitals (e.g., $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$). In particular, we denote by $\mathbf{V} = \{1, \ldots, n\}$ the set of all n variables in the domain. We name the variables by their indices in $\mathbf{V}$; for instance, we refer to the third variable in $\mathbf{V}$ simply by 3. We assume that all variables in the domain are discrete. We denote the data set by D and its size (number of data points) by N. We use the notation $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})$ to denote that $\mathbf{X}$ is independent of $\mathbf{Y}$ conditioned on $\mathbf{Z}$, for disjoint sets of variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, while $(\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})$ denotes conditional dependence. For the sake of readability, we will slightly abuse this notation and use $(X \perp\!\!\!\perp Y \mid \mathbf{Z})$ as shorthand for $(\{X\} \perp\!\!\!\perp \{Y\} \mid \mathbf{Z})$.

A Bayesian network (**BN**) is a directed graphical model which represents the joint probability distribution over $\mathbf{V}$. Each node in the graph represents one of the random variables in the domain. The structure of the network represents a set of conditional independences on the domain

$$
\begin{array}{rl}
\textbf{(Symmetry)} & (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}) \iff (\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}) \\
\textbf{(Decomposition)} & (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}) \wedge (\mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z}) \\
\textbf{(Weak Union)} & (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}) \\
\textbf{(Contraction)} & (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}) \wedge (\mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}) \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) \\
\textbf{(Intersection)} & (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}) \wedge (\mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}) \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z})
\end{array}
\tag{5}
$$

$$
\begin{array}{rl}
\textbf{(Symmetry)} & (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}) \iff (\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}) \\
\textbf{(Decomposition)} & (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}) \wedge (\mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z}) \\
\textbf{(Intersection)} & (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}) \wedge (\mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}) \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) \\
\textbf{(Weak Union)} & (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}) \\
\textbf{(Contraction)} & (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}) \wedge (\mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y}) \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}) \\
\textbf{(Weak Transitivity)} & (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}) \wedge (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \cup \gamma) \implies (\mathbf{X} \perp\!\!\!\perp \gamma \mid \mathbf{Z}) \vee (\gamma \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}) \\
\textbf{(Chordality)} & (\alpha \perp\!\!\!\perp \beta \mid \gamma \cup \delta) \wedge (\gamma \perp\!\!\!\perp \delta \mid \alpha \cup \beta) \implies (\alpha \perp\!\!\!\perp \beta \mid \gamma) \vee (\alpha \perp\!\!\!\perp \beta \mid \delta)
\end{array}
\tag{6}
$$

variables. Given the structure of a BN, the set of independencies implied by it can be identified by a process called *d-separation*: All independencies identified by d-separation are implied by the model structure. If in addition all remaining triplets $(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})$ correspond to dependencies, we say that the BN is *directed graph-isomorph*, abbreviated *DAG-isomorph*, or simply *causal*. The concept of isomorphism is closely related to faithfulness. A graph G is said to be *faithful* to some distribution if exactly those independences that exist in the distribution and no others are returned by d-separation on G. In this paper we assume faithfulness. We also make the assumption of *causal sufficiency*. A domain is causally sufficient if there exist no *hidden* or *latent* variables in it.

As mentioned above, independence-based algorithms operate by conducting a series of conditional independence queries. For these we assume that an *independence-query oracle* exists that is able to provide such information. This approach can be viewed as an instance of a statistical query oracle (Kearns and Vazirani, 1994). In practice such an oracle does not exist, but is frequently implemented approximately by a statistical test evaluated on the data set (for example, this can be Pearson's conditional independence χ^2 (chi-square)test (Agresti, 2002), Wilk's G^2 test, a mutual information test etc.). In this work we used Wilk's G^2 test (Agresti, 2002). To determine conditional independence between two variables X and Y given a set $\mathbf{Z}$ from data, the statistical test G^2 (and any other independence test based on hypothesis testing, e.g., the $|chi^2$ test) returns a *p-value*, which is the probability of error in assuming that the two variables are dependent when in fact they are not. If the p-value of a test is $p(X, Y \mid \mathbf{Z})$, the statistical test concludes independence if and only if $1 - p(X, Y \mid \mathbf{Z})$ is smaller than or equal to a confidence threshold α i.e.,

$$(X \perp\!\!\!\perp Y \mid \mathbf{Z}) \iff p(X, Y \mid \mathbf{Z}) \geq 1 - \alpha. \tag{7}$$

Common values for α are 0.95, 0.99, and 0.90.

The conditional independences and dependences of a domain are connected through a set of general rules. Let us imagine a meta-space of binary variables, each corresponding to the truth value of the independence of a triplet $(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})$ (e.g., `true` for independence and `false` for dependence). Each point in this space corresponds to a conditional independence assignment to all possible triplets in the domain. In this conceptual space not all points are tenable; in particular a set of rules exists, presented in Pearl (1988) and shown in Eqs. (5), that constrain the truth values of independencies corresponding to triplets. For domains for which there exists a faithful Bayesian network a more relaxed set of properties hold, shown in Eqs. (6), where α, β, γ and δ correspond to single variables. In both sets of axioms, Intersection holds if the probability distribution of the

domain is positive i.e., every assignment to all variables in the domain has a non-zero probability.

In the next section we describe the argumentation framework in general, followed by its application to our problem of answering independence queries from knowledge bases that contain sets of potentially inconsistent independence propositions.

3 The Argumentation Framework

As we mentioned previously, we model the framework of learning a causal model through independence queries as a set of rules (Eqs. (5) or (6)) and a knowledge base (**KB**) that contains independence propositions that may be inconsistent.

There exist two major approaches for reasoning with inconsistent knowledge that correspond to two different attitudes: One is to resolve the inconsistencies by removing a subset of propositions such that the resulting KB becomes consistent; this is called *belief revision* in the literature (Gärdenforst, 1992; Gärdenforst and Rott, 1995; Shapiro, 1998; Martins, 1992). A known shortcoming of belief revision (Shapiro, 1998) stems from the fact that it removes propositions, which, besides discarding potentially valuable information, has the same potential problem as the problem that we are trying to solve: an erroneous modification of the KB may have unintended negative consequences if later more propositions are inserted in the KB. A second approach to inconsistent KBs is to allow inconsistencies but uses rules that may be possibly contained in it to deduce which truth value of a proposition query is "preferred" in some way. One instance of this approach is *argumentation* (Dung, 1995; Loui, 1987; Prakken, 1997; Prakken and Vreeswijk, 2002), a sound approach that allows inconsistencies but uses a proof procedure that is able to deduce (if possible) that one of the truth values of certain propositions is preferred over its negation; this may happen because the latter is contradicted by other rules and/or propositions in the KB (a more precise definition is given below). Argumentation is a reasoning model that belongs to the broader class of defeasible logics (Pollock, 1992; Prakken, 1997). Our approach uses the argumentation framework of Amgoud and Cayrol (2002) that considers preferences over arguments, extending Dung's more fundamental framework (Dung, 1995). Preference relations give an extra level of specificity for comparing arguments, allowing a more refined form of selection between conflicting propositions. Preference-based argumentation is presented in more detail in the Section 3.2.

We proceed now to describe the argumentation framework.

Definition 1. *An* argumentation framework *is a pair $\langle \mathcal{A}, \mathcal{R} \rangle$, where $\mathcal{A}$ is a set of arguments and $\mathcal{R}$ is a binary relation representing a defeasibility relationship between arguments, i.e., $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$. $(A, B) \in \mathcal{R}$ or equivalently "$A\ \mathcal{R}\ B$" means that argument A defeats the argument B. We also say that A and B are in conflict.*

An example of the defeat relation $\mathcal{R}$ is *logical defeat*, which occurs when an argument contradicts another logically.

The elements of the argumentation framework are not propositions but *arguments*. Given an inconsistent knowledge base $\mathcal{K} = \langle \Sigma, \Psi \rangle$ with a set of propositions Σ and a set of inference rules Ψ, arguments are defined formally as follows.

Definition 2. *An* argument *over knowledge base $\langle \Sigma, \Psi \rangle$ is a pair (H, h) where $H \subseteq \Sigma$ such that:*

- *H is consistent,*
- *$H \vdash_\Psi h$,*
- *H is minimal (with respect to set inclusion).*

H is called the support *and h the* conclusion *or* head *of the argument.*

In the above definition $\vdash_\Psi$ stands for classical inference over the set of inference rules Ψ. Intuitively an argument (H, h) can be thought as an "if-then" rule i.e., "if H then h". In inconsistent knowledge bases two arguments may contradict or *defeat* each other. The defeat relation is defined through the *rebut* and *undercut* relations, defined below.

Definition 3. *Let (H_1, h_1), (H_2, h_2) be two arguments.*

- (H_1, h_1) rebuts (H_2, h_2) *iff* $h_1 \equiv \neg h_2$.
- (H_1, h_1) undercuts (H_2, h_2) *iff* $\exists h \in H_2$ *such that* $h \equiv \neg h_1$.

(The symbol "$\equiv$" stands for logical equivalence.) In other words, $(H_1, h_1)\ \mathcal{R}\ (H_2, h_2)$ if and only if (H_1, h_1) either *rebuts* or *undercuts* (H_2, h_2).

The objective of argumentation is to decide on the acceptability of each argument. There are three possibilities: an argument can be accepted, rejected, or neither. This partitions the space of arguments $\mathcal{A}$ in three classes:

- The class $Acc_\mathcal{R}$ of *acceptable arguments*. Intuitively, these are the "good" arguments. In the case of an inconsistent knowledge base, these will be inferred from the base.
- The class $Rej_\mathcal{R}$ of *rejected arguments*. These are the arguments defeated by acceptable arguments. When applied to an inconsistent knowledge base, these will not be inferred from it.
- The class $Ab_\mathcal{R}$ of arguments *in abeyance*. These arguments are neither acceptable nor rejected.

The semantics of acceptability proposed by Dung dictates that an argument should be accepted if it is not defeated, or if it is defended by acceptable arguments i.e., each of its defeaters is itself defeated by an acceptable argument. This is formalized in the following definitions.

Definition 4. *Let $\langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation framework, and $S \subseteq \mathcal{A}$. An argument A is* defended *by S if and only if $\forall B$, if $B\ \mathcal{R}\ A$ then $\exists C \in S$ such that $C\ \mathcal{R}\ B$.*

Dung characterizes the set of acceptable arguments by a monotonic function $\mathcal{F}$, i.e., $\mathcal{F}(S) \subseteq \mathcal{F}(S \cup T)$ for some S and T. Given a set of arguments $S \subseteq \mathcal{A}$ as input, $\mathcal{F}$ returns the set of all arguments defended by S:

Definition 5. *Let $S \subseteq \mathcal{A}$. Then $\mathcal{F}(S) = \{A \in \mathcal{A} \mid A$ is defended by $S\}$.*

Slightly overloading our notation, we define $\mathcal{F}(\varnothing)$ to contain the set of arguments that are not defeated, i.e., defend themselves.

Definition 6. *Let $\langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation framework, and let $A \in \mathcal{A}$ be some argument. We say A* defends itself *if it is not defeated by any other argument, i.e. $\forall B \neq A \in \mathcal{A}$, $\neg(B\ \mathcal{R}\ A)$.*

Definition 7. *$\mathcal{F}(\varnothing) = \{A \in \mathcal{A} \mid A$ defends itself $\}$.*

Dung proved that the set of acceptable arguments is the least fix-point of $\mathcal{F}$, i.e., the smallest set S such that $\mathcal{F}(S) = S$.

Proposition 1. *Let $\langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation framework. The set of acceptable arguments $Acc_\mathcal{R}$ is the least fix-point of the function $\mathcal{F}$.*

Dung also showed that if the argumentation framework $\langle \mathcal{A}, \mathcal{R} \rangle$ is finitary i.e., for each argument A there are finitely many arguments that defeat A, the least fix-point of function $\mathcal{F}$ can be obtained by iterative application of $\mathcal{F}$ to the empty set. We can understand this intuitively: From our semantics of acceptability it follows that all arguments in $\mathcal{F}(\varnothing)$ are accepted. Also, every argument in $\mathcal{F}(\mathcal{F}(\varnothing))$ must be acceptable as well since each of its arguments is defended by acceptable arguments. This reasoning can be applied recursively until a fix-point is reached. The fix-point S is the set of arguments that cannot defend any other argument not in S i.e., no other argument is accepted. This suggests a simple algorithm for computing the set of acceptable arguments. Algorithm 1 shows a recursive procedure for this, based on the above definition. The algorithm takes as input an argumentation framework $\langle \mathcal{A}, \mathcal{R} \rangle$ and the set S of arguments found acceptable so far i.e., $S = \varnothing$.

Algorithm 1 Recursive computation of acceptable arguments: $Acc_{\mathcal{R}} = \mathcal{F}(\mathcal{A}, \mathcal{R}, S)$

```
S' ⟵ S ∪ {A ∈ 𝒜 | A is defended by S}
if S = S' then
    return S'
else
    return ℱ(𝒜, ℛ, S')
```

Let us illustrate these ideas with an example.

Example 2. *Let $\langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation framework defined by $\mathcal{A} = \{A, B, C\}$ and $\mathcal{R} = \{(A, B), (B, C)\}$. The only argument that is not defeated (i.e., defends itself) is A, and therefore $\mathcal{F}(\varnothing) = \{A\}$. Argument B is defeated by the acceptable argument A, so B cannot be defended and is therefore rejected i.e., $B \in Rej_{\mathcal{R}}$. Argument C, though defeated by B, is defended by (acceptable argument) A which defeats B, so C is acceptable. The set of acceptable arguments is therefore $Acc_{\mathcal{R}} = \{A, C\}$ and the set of rejected arguments is $Rej_{\mathcal{R}} = \{B\}$.*

3.1 Argumentation in Independence Knowledge Bases

We can apply the argumentation framework to our problem of answering queries from knowledge bases that contain a number of potentially inconsistent independencies and dependencies and a set of rules that express relations among them.

Definition 8. *An* independence knowledge base *(**IKB**) is a knowledge base $\langle \Sigma, \Psi \rangle$ such that its proposition set Σ contains only independence propositions of the form $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})$ or $(\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})$, and its inference rules Ψ are either the set of axioms shown in Eqs. (5), in which case we call it a* general independence knowledge base*, or the set of axioms shown in Eqs. (6), in which case we call it a* specific *or* causal independence knowledge base*.*

For IKBs, the set of arguments $\mathcal{A}$ is constructed in two steps. First, for each proposition $\sigma \in \Sigma$ (independence or dependence) we add to $\mathcal{A}$ the argument $(\{\sigma\}, \sigma)$. This is a valid argument according to Definition 2 since its support $\{\sigma\}$ is (trivially) consistent, it (trivially) implies the head σ, and it is minimal (the pair $(\{\varnothing\}, h)$ is not a valid argument since $\varnothing$ is equivalent to the proposition `true` which does not entail h). Arguments of the form $(\{\sigma\}, \sigma)$ are called *propositional arguments* since they correspond to single propositions. The second step in the construction of the set of arguments $\mathcal{A}$ concerns rules and proceeds as follows: for each inference rule $(\Phi_1 \wedge \Phi_2 \ldots \wedge \Phi_n \implies \Phi) \in \Psi$, and each subset of Σ that matches exactly the set of antecedents, i.e., each

subset $\{\varphi_1 \wedge \varphi_2 \ldots \wedge \varphi_n\}$ of Σ such that $\Phi_1 \equiv \varphi_1, \Phi_2 \equiv \varphi_2 \ldots \Phi_n \equiv \varphi_n$, we add argument $(\{\varphi_1 \wedge \varphi_2 \ldots \varphi_n\}, \varphi)$ to $\mathcal{A}$.[1]

IKBs can be augmented with a set of preferences that allows one to take into account the reliability of tests when deciding on the truth value of independence queries. This is described in the next section.

3.2 Preference-based Argumentation Framework

Following Amgoud and Cayrol (2002), we now refine the argumentation framework of Dung (1995) for cases where it is possible to define a preference order Π over arguments.

Definition 9. *A* preference-based argumentation framework ***(PAF)*** *is a triplet* $\langle \mathcal{A}, \mathcal{R}, \Pi \rangle$ *where* $\mathcal{A}$ *is a set of arguments,* $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ *is a binary relation representing a defeat relationship between pairs of arguments, and* Π *is a (partial or complete) ordering over* $\mathcal{A} \times \mathcal{A}$.

For the case of inconsistent knowledge bases, preference Π over arguments follows the preference π over their support i.e., stronger support implies a stronger argument, which is given as a partial or total order over sets of propositions. Formally:

Definition 10. *Let* $\mathcal{K} = \langle \Sigma, \Psi \rangle$ *be a knowledge base,* π *be a (partial or total) ordering on subsets of* Σ *and* (H, h), (H', h') *two arguments over* $\mathcal{K}$. *Argument* (H, h) *is* π-preferred *to* (H', h') *(denoted* $(H, h) \gg_\pi (H', h')$*) if and only if* H *is preferred to* H' *with respect to* π.

In what follows we overload our notation by using π to denote either the ordering over arguments or over their supports.

The defeat and preference relations can be combined into a refined defeat relation called *attack*.

Definition 11. *Let* $\langle \mathcal{A}, \mathcal{R}, \pi \rangle$ *be a PAF, and* A, $B \in \mathcal{A}$ *be two arguments. We say* B attacks A *if and only if* $B \mathcal{R} A$ *and* $\neg(A \gg_\pi B)$.

We can see that a preference-based argumentation framework is a special case of the more general argumentation framework, having a more refined defeat relation. Therefore the same conclusions apply, in particular Proposition 1, which allows us to compute the set of acceptable arguments of a PAF using Alg. 1.

We can now apply these ideas to construct a more reliable approximation to the independence-query oracle.

3.3 Preference-based Argumentation in Independence Knowledge Bases

In this section we describe how to apply the preference-based argumentation framework of Section 3.2 to improve the reliability of conditional independence tests conducted on (possibly small) data sets.

A preference-based argumentation framework has three components. The first two, namely $\mathcal{A}$ and $\mathcal{R}$ are identical to general argumentation frameworks. We now describe how we construct the third component, namely the preference ordering π over subsets H of Σ, in IKBs. We define it using the probability $\nu(H)$ that all propositions in H are correct, that is

$$H \gg_\pi H' \Longleftrightarrow \nu(H) \geq \nu(H').$$

[1] This is equivalent to propositionalizing the set of rules, some of which may be first-order (the rules of Eqs. (5) and (6) are universally quantified over all sets of variables). As this may be expensive (exponential in the number of propositions), in practice it may not be implemented in this way, instead matching appropriate rules on the fly during the argumentation inference process.

We compute the probability $\nu(H)$ by assuming independence among the propositions. Overloading notation and denoting by $\nu(h)$ the probability of an individual proposition h being correct, the probability of all elements in H being correct under this assumption of independence is

$$\nu(H) = \prod_{h \in H} \nu(h). \tag{8}$$

In our case we have independence propositions. The probability that an independence proposition is correct can be computed in different ways, depending on the particular choice of independence oracle chosen. In this work we use Wilk's G^2 test. As discussed in Section 2, the p-value $p(X, Y \mid \mathbf{Z})$ computed by this test is the probability of error in assuming that X and Y are dependent when in fact they are not. Therefore, the probability of a test returning dependence of being correct is

$$\nu_D(X \not\perp\!\!\!\perp Y \mid \mathbf{Z}) = 1 - p(X, Y \mid \mathbf{Z}) \tag{9}$$

where the subscript D indicates that this expression is valid only for dependencies.

The probability of correctly reporting an independence is defined in terms of the β-value, the probability of incorrectly reporting independence when in fact the variables are dependent:

$$\nu_I(X \perp\!\!\!\perp Y \mid \mathbf{Z}) = 1 - \beta(X, Y \mid \mathbf{Z}) \tag{10}$$

where again the subscript I indicates that it is valid only for independences.

To the best of our knowledge, the general computation of the β-value is an open problem. While computing the p-value involves evaluating the probability of a statistic under the distribution generated by the independence model, i.e., a model under which the variables are independent, which for discrete domains is unique, computing β is difficult because there are infinitely many possible models in which the variables are dependent. In statistical applications, the β value is commonly approximated by assuming one particular dependence model if some prior knowledge is available. In the absence of such information however we take an alternative approach of approximating the β-value from the p-value. We estimate the β-value of a test on triplet $(X, Y \mid \mathbf{Z})$ from the p-value assuming the following heuristic constraints on β:

$$\beta(p(X, Y \mid \mathbf{Z})) = \begin{cases} \frac{1}{2+|\mathbf{Z}|} & \text{if } p(X, Y \mid \mathbf{Z}) = 1 \\ \alpha + \frac{1}{2+|\mathbf{Z}|} & \text{if } p(X, Y \mid \mathbf{Z}) = 0 \\ 1 - \alpha & \text{if } p(X, Y \mid \mathbf{Z}) = 1 - \alpha \end{cases}$$

The first constraint (for $p(X, Y \mid \mathbf{Z}) = 1$) is justified by the intuition that when the p-value of the test is close to 1, the test statistic is close to its value under the model that assumes independence, and thus we would give more preference to the "independence" decision. The situation for the second case ($p(X, Y \mid \mathbf{Z}) = 0$) is reversed—the statistic is very far from the expected one under independence, and therefore independence is not preferred. Both values are tempered by the number of variables in the conditioning set. This reflects the practical consideration that, as the number of variables involved in the test $2 + |\mathbf{Z}|$ increases, given a fixed data set, the reliability of the test diminishes, going to 0 as $|\mathbf{Z}| \to \infty$; in the limit therefore the preference of an independence test becomes a horizontal line, crossing the vertical axis at $\beta = \alpha$. The third assumption is related to fairness: In the absence of non-propositional arguments (i.e., in the absence of inference rules in the knowledge-base), the independence decisions of the argumentation framework should match those of the purely statistical tests. Otherwise, changes in the outcome of tests may be due to simply bias in the independence decision that favors dependence or independence i.e., it is equivalent to

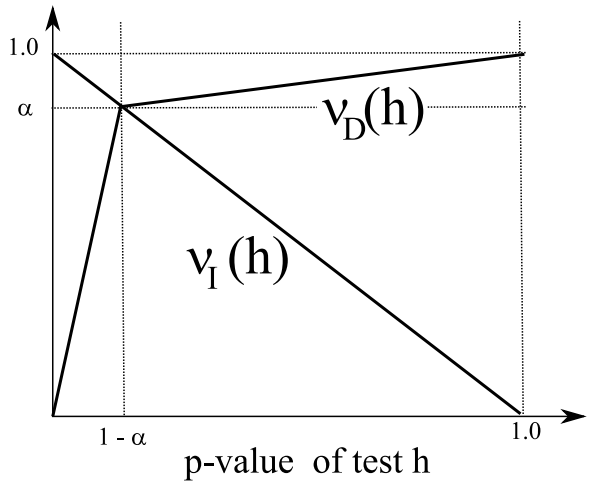

Figure 1: The probability of correct independence $\nu_I(h) = 1-\beta(p(h))$ and the probability of correct dependence $\nu_D(h) = 1 - p(h)$ as a function of the p-value $p(h)$ of test h.

an arbitrary change to the threshold of the statistical test, and the comparison of the two tests would not be a fair one.

The remaining values of β are approximated by linear interpolation among the above points. The result is summarized in Fig. (1), which shows the probabilities of dependence ν_D (i.e., $1 - p$) and ν_I (i.e., $1 - \beta$) versus p.

We now use the following example to illustrate how preference-based argumentation can be used to resolve the inconsistencies of Example 1.

Example 3. *Let us extend the IKB of Example 1 with the following preference values for its propositions and rules.*

$$\begin{aligned} \mathit{Pref}\left[(\{0\} \perp\!\!\!\perp \{1\} \mid \{2,3\})\right] &= 0.8 \\ \mathit{Pref}\left[(\{0\} \perp\!\!\!\perp \{4\} \mid \{2,3\})\right] &= 0.7 \\ \mathit{Pref}\left[(\{0\} \not\perp\!\!\!\perp \{1,4\} \mid \{2,3\})\right] &= 0.5 \end{aligned}$$

Following the IKB construction procedure described in the previous section, the above propositions correspond to the following arguments, respectively:

$$\left(\left\{(0 \perp\!\!\!\perp 1 \mid \{2,3\})\right\}, (0 \perp\!\!\!\perp 1 \mid \{2,3\})\right) \tag{11}$$

$$\left(\left\{(0 \perp\!\!\!\perp 4 \mid \{2,3\})\right\}, (0 \perp\!\!\!\perp 4 \mid \{2,3\})\right) \tag{12}$$

$$\left(\left\{(0 \not\perp\!\!\!\perp \{1,4\} \mid \{2,3\})\right\}, (0 \not\perp\!\!\!\perp \{1,4\} \mid \{2,3\})\right) \tag{13}$$

and rule (4) corresponds to the following argument

$$\left(\left\{(0 \perp\!\!\!\perp 1 \mid \{2,3\}), (0 \perp\!\!\!\perp 4 \mid \{2,3\})\right\}, (0 \perp\!\!\!\perp \{1,4\} \mid \{2,3\})\right). \tag{14}$$

The preference of each argument $(\{\sigma\}, \sigma)$ is equal to the preference value of $\{\sigma\}$, according to Definition 10, which, as it contains only a single proposition, is equal to the preference of σ.

Therefore,

$$\begin{aligned} \mathit{Pref}\left[\left(\left\{(\{0\}\perp\!\!\!\perp\{1\} \mid \{2,3\})\right\}, (\{0\}\perp\!\!\!\perp\{1\} \mid \{2,3\})\right)\right] &= 0.8 \\ \mathit{Pref}\left[\left(\left\{(\{0\}\perp\!\!\!\perp\{4\} \mid \{2,3\})\right\}, (\{0\}\perp\!\!\!\perp\{4\} \mid \{2,3\})\right)\right] &= 0.7 \\ \mathit{Pref}\left[\left(\left\{(\{0\}\not\perp\!\!\!\perp\{1,4\} \mid \{2,3\})\right\}, (\{0\}\not\perp\!\!\!\perp\{1,4\} \mid \{2,3\})\right)\right] &= 0.5. \end{aligned}$$

The preference of argument (14) equals the preference of the set of its antecedents, which, according to Eq. (8), is equal to the product of their individual preferences i.e.,

$$\mathit{Pref}\left[\left(\left\{(0\perp\!\!\!\perp 1 \mid \{2,3\}), (0\perp\!\!\!\perp 4 \mid \{2,3\})\right\}, (0\perp\!\!\!\perp 1 \mid \{2,3\})\right)\right] = 0.8 \times 0.7 = 0.56.$$

We now show how argumentation resolves the inconsistency between proposition (3) and rule (4) of Example 1. Even though proposition (3) and rule (4) contradict each other logically, i.e., their corresponding arguments (13) and (14) defeat each other, argument (14) defends itself because its preference is 0.56 *which is larger than* 0.5*, the preference of its defeater argument (13). Also, since no other argument defeats (14), it is acceptable, and (13), being attacked by an acceptable argument, must be rejected. We therefore see that using preferences the inconsistency of Example 1 has been resolved in favor of rule (4).*

We now extend Example 3 to illustrate the defend relation, i.e., how an argument can be defended by some other argument. The example also illustrate an alternative resolution for the inconsistency of Example 1, this time in favor of proposition (3).

Example 4. *Let us extend the IKB of Example 3 with two additional independence propositions and an additional rule.*

The new propositions and their corresponding preferences are:

$$\begin{aligned} \mathit{Pref}\left[(0\perp\!\!\!\perp 5 \mid \{2,3\})\right] &= 0.8 \\ \mathit{Pref}\left[(0\not\perp\!\!\!\perp\{1,5\} \mid \{2,3\})\right] &= 0.9. \end{aligned}$$

and the new rule is:

$$(0\perp\!\!\!\perp 5 \mid \{2,3\}) \wedge (0\not\perp\!\!\!\perp\{1,5\} \mid \{2,3\}) \implies (0\not\perp\!\!\!\perp 1 \mid \{2,3\}).$$

This rule is an instance of the Composition axiom in contrapositive form.

The corresponding arguments are therefore:

$$\begin{aligned} \mathit{Pref}\left[\left(\left\{(\{0\}\perp\!\!\!\perp\{5\} \mid \{2,3\})\right\}, (\{0\}\perp\!\!\!\perp\{5\} \mid \{2,3\})\right)\right] &= 0.8 \\ \mathit{Pref}\left[\left(\left\{(\{0\}\not\perp\!\!\!\perp\{1,5\} \mid \{2,3\})\right\}, (\{0\}\not\perp\!\!\!\perp\{1,5\} \mid \{2,3\})\right)\right] &= 0.9 \end{aligned}$$

corresponding to the two propositions, and

$$\mathit{Pref}\left[\left(\left\{(0\perp\!\!\!\perp 5 \mid \{2,3\}), (0\not\perp\!\!\!\perp\{1,5\} \mid \{2,3\})\right\}, (0\not\perp\!\!\!\perp 1 \mid \{2,3\})\right)\right] = 0.8 \times 0.9 = 0.72 \qquad (15)$$

corresponding to the rule.

As in Example 3, argument (13) is attacked by argument (14). If the IKB was as in Example 3, (14) would had been acceptable and (13) would have been rejected. However, the additional argument (15) defeats (undercuts) (14), by logically contradicting its antecedent $(\{0\}\perp\!\!\!\perp\{1\} \mid \{2,3\})$*. Since (15) also attacks (14) i.e., its preference* 0.72 *is larger than* 0.56*, the preference of (14), (15) defends all arguments that are attacked by argument (14), in particular (13). Note this is not sufficient for accepting (13) as it has not been proved that its defender (15) is itself acceptable. We leave the proof of this as an exercise for the reader.*

3.4 Argumentative independence tests

We are now ready to present our argumentation-based independence test (**AIT**). Given an input triplet $(X, Y \mid \mathbf{Z})$ and a preference-based argumentation framework, an AIT responds independence $(X \perp\!\!\!\perp Y \mid \mathbf{Z})$ or dependence $(X \not\perp\!\!\!\perp Y \mid \mathbf{Z})$ of X and Y given $\mathbf{Z}$ by applying the framework to determine the acceptability of their corresponding propositional arguments i.e. it responds

$$\begin{aligned} &(X \not\perp\!\!\!\perp Y \mid \mathbf{Z}) \qquad \text{if argument } (\{(X \not\perp\!\!\!\perp Y \mid \mathbf{Z})\}, (X \not\perp\!\!\!\perp Y \mid \mathbf{Z})) \text{ is accepted, or} \\ &(X \perp\!\!\!\perp Y \mid \mathbf{Z}) \qquad \text{if argument } (\{(X \perp\!\!\!\perp Y \mid \mathbf{Z})\}, (X \perp\!\!\!\perp Y \mid \mathbf{Z})) \text{ is accepted.} \end{aligned} \tag{16}$$

Although we have not observed it in practice, in the case that both propositional arguments are accepted or both are not accepted i.e., each of the propositional arguments is either rejected or in abeyance, we simply respond $(1 - p(X, Y \mid \mathbf{Z}) \leq \alpha)$ i.e., the independence value of the statistical independence test.

The rationale behind the AIT is that a propositional argument $(\{\sigma\}, \sigma)$ contains only the head σ in its support and therefore attacks to the argument are attacks to σ: The argument is defeated (rebutted or undercut) if and only if σ is contradicted, and its preference is lower than the preference π of another argument if and only if the preference of σ is lower than π. The semantics of acceptability therefore propagates to σ.

We now illustrate the use of AIT with an extension of Example 3.

Example 5. *Let us consider an extension of Example 3 to illustrate the use of the AIT to decide on the independence or dependence of input triplet* $(\{0\}, \{1, 4\} \mid \{2, 3\})$*. According to Eq. (16) the decision depends on the status of the two propositional arguments:*

$$(\{(\{0\} \perp\!\!\!\perp \{1, 4\} \mid \{2, 3\})\}, (\{0\} \perp\!\!\!\perp \{1, 4\} \mid \{2, 3\})), \text{ and} \tag{17}$$

$$(\{(\{0\} \not\perp\!\!\!\perp \{1, 4\} \mid \{2, 3\})\}, (\{0\} \not\perp\!\!\!\perp \{1, 4\} \mid \{2, 3\})) \tag{18}$$

Argument (18) is equivalent to argument (13) of Example 3, that was proven to be rejected. According to Eq.(16), the AIT therefore does not decide dependence.

To query the acceptance of argument (17) we add it to the arguments set of the argumentation framework of Example 3 assuming the following preference value

$$\textit{Pref}\,[(\{(\{0\} \perp\!\!\!\perp \{1, 4\} \mid \{2, 3\})\}, (\{0\} \perp\!\!\!\perp \{1, 4\} \mid \{2, 3\}))] = 0.92. \tag{19}$$

With this preference value, propositional argument (17) is not attacked by its unique defeater (13). Therefore, it must be accepted and according to Eq.(16), the independence $(\{0\} \perp\!\!\!\perp \{1, 4\} \mid \{2, 3\})$ *is inferred for triplet* $(\{0\}, \{1, 4\} \mid \{2, 3\})$ *in this IKB.*

4 Experimental Results

As our main focus in the present paper was to demonstrate that the argumentation approach does indeed improve the accuracy of independence tests on small data sets, we did not focus on issues of efficiency in our experimental evaluation. As such we generated our set of propositional arguments i.e., arguments of the form $(\{\sigma\}, \sigma)$, by iterating over all possible triplets $(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})$ and inserting them in the knowledge base, together with their preference, as described in Section 3.1. Similarly, for the set of axioms that we used in each case i.e., either Eq. (5) or Eq. (6), we iterated over all possible matches of each rule, inserting the corresponding instantiated rule in the knowledge base together with its preference, again as described in Section 3.1. The reason for including all propositional and

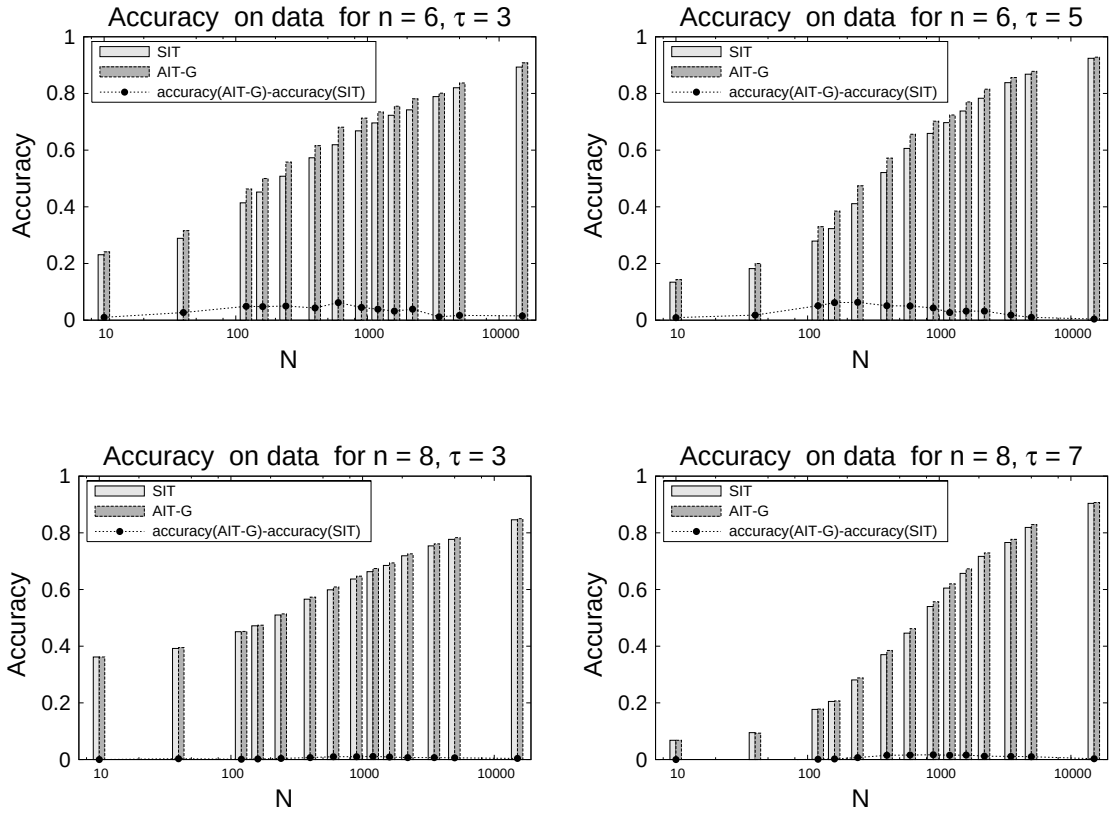

Figure 2: Comparison of statistical tests (SIT) vs. argumentative tests on the general axioms (AIT-G) for domain size $n = 3$ and $\tau = 3, 5$ and for $n = 8$ and $\tau = 3, 7$. The histograms show the absolute value of the accuracy while the line curves shows their difference i.e., a positive value corresponds to an improvement in the accuracy of AIT-G over the accuracy of SIT.

rule-based arguments in our IKB is to allow the argumentation framework to consider all possible arguments in favor or against an independence query. Also, our implementation uses Alg. 1 for inference while answering a query from our preference-based IKB. The time complexity of algorithm Alg. 1 is linear with the size $|\mathcal{A}|$ of the space of arguments, but $|\mathcal{A}|$ itself grows super-exponentially with the domain size n. This prevented us from exploring domain sizes larger than $n = 8$. Clearly the present implementation is suboptimal, but our (improved) accuracy results demonstrate the utility of our approach. The design of a more efficient algorithm for inference is a useful direction of future research. This is briefly described in Section 5.

We conducted experiments on sampled and real-world data sets and compared the performance of the argumentative independence tests (**AIT**s) versus their statistical counterpart (**SIT**s) for varying reliability conditions (obtained by conducting experiments on varying data set sizes). We measured the performance of each independence test (SIT or AIT) by its accuracy. The accuracy was estimated by performing a number of conditional independence tests on data, and comparing the result (`true` or `false`) of each of these with the true value of the corresponding independence, computed by querying the underlying model for the conditional independence value of the same test. This approach is similar to estimating accuracy in a classification task over unseen instances but with inputs here being triplets $(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})$ and the class attribute being the value of the corresponding conditional independence test.

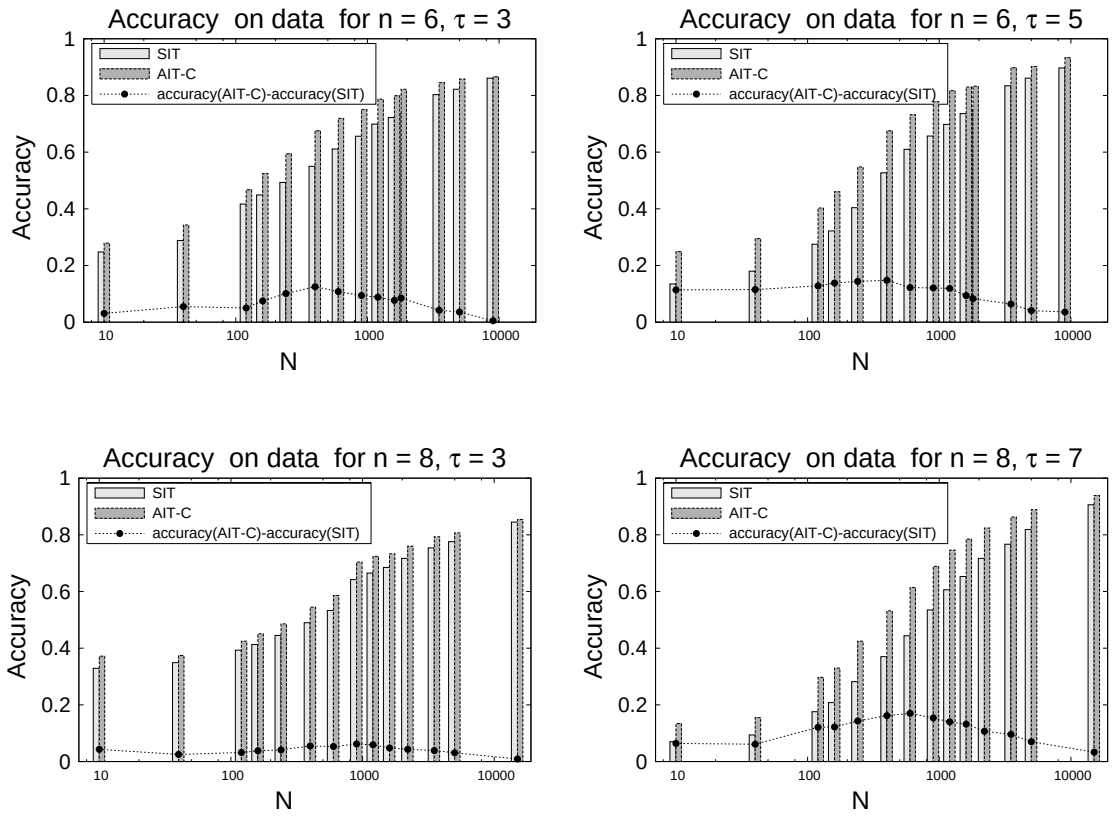

Figure 3: Comparison of statistical tests (SIT) vs. argumentative tests on the causal axioms (AIT-C) for domain size $n = 3$, maximum degree $\tau = 3, 5$ and domain size $n = 8$, maximum degree $\tau = 3, 7$. The histogram shows the absolute value of the accuracies and the line curve shows their difference i.e., a positive value correspond to an improvement in the accuracy of AIT-C over the accuracy of SIT.

In the next section we present results for data sampled from Bayesian networks, where the underlying model is known and can be queried for conditional independence using d-separation. Following this, we present results of real-wold data experiments where the underlying model is unknown and thus the true values of the independences must be approximated; this is explained in detail below.

4.1 Sampled Data Experiments

In this set of experiments we compare the accuracy of argumentative tests (AITs) versus purely statistical tests (SITs) on several data sets sampled from a number of randomly generated Bayesian networks. Sampled data experiments have the advantage of a more precise estimation of the accuracy since the underlying model is known. We present experiments for two versions of the argumentative test, one that uses Pearl's general axioms of Eq. (5), denoted **AIT-G**, and another that uses Pearl's causal axioms of Eq. (6), denoted **AIT-C**.

The data was sampled from randomly generated Bayesian networks of different number of nodes n and different maximum degrees per node τ (corresponding to different arc densities in the resulting graphs) using *BNGenerator* (Ide et al., 2002), a publicly available Java package. For $n = 6$ we generated ten networks with $\tau = 3$ and ten networks with $\tau = 5$. For $n = 8$ we generated

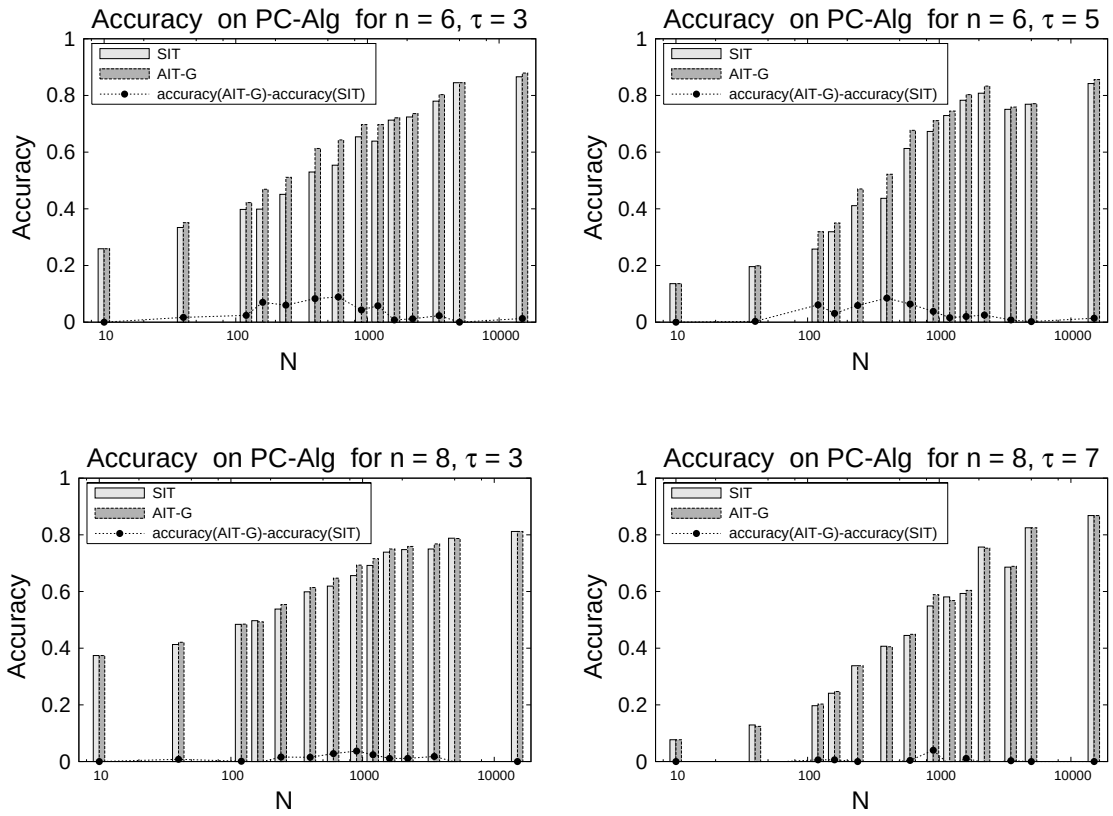

Figure 4: Comparison of the output of PC-Algorithm for SIT and AIT-G for domain size $n = 3$, maximum degree $\tau = 3, 5$ and domain size $n = 8$, maximum degree $\tau = 3, 7$. The histogram shows the absolute value of the accuracies and the line curve shows their difference i.e., a positive value correspond to an improvement in the accuracy of AIT-G over the accuracy of SIT.

ten networks for $\tau = 3$ and another ten for $\tau = 7$. For each data set D in these four groups, we conducted experiments on subsets of D containing an increasing number of data points. This was done in order to assess the performance of the independence tests (SITs or AITs) on varying conditions of reliability, as the reliability of a test typically decreases with decreasing data set size.

For each experiment we report the estimated accuracy, calculated by comparing the result of a number of conditional independence tests (SITs or AITs) on data with the true value of independence, computed by querying the underlying model for the conditional independence of the same test using d-separation. Since the number of possible tests is exponential, we estimated the independence accuracy by sampling 2,000 triplets $(X, Y, \mathbf{Z})$ randomly, evenly distributed among all possible conditioning set sizes $m \in \{0, \ldots, n-2\}$ (i.e., $2,000/(n-1)$ tests for each m). Denoting by $\mathcal{T}$ this set of 2,000 triplets, by $t \in \mathcal{T}$ a triplet, by $I_{\text{true}}(t)$ the result of a test performed on the underlying model, and by $I_{\text{data-}\mathcal{Y}}(t)$ the results of performing a test of type $\mathcal{Y}$ on data, for $\mathcal{Y}$ equal to SIT, AIT-G or AIT-C, the estimated accuracy of test type $\mathcal{Y}$ is defined as

$$\widehat{accuracy}_{\mathcal{Y}}^{\text{data}} = \frac{1}{|\mathcal{T}|}\left|\left\{t \in \mathcal{T} \mid I_{\text{data-}\mathcal{Y}}(t) = I_{\text{true}}(t)\right\}\right|.$$

Figure 2 shows a comparison of the argumentative test AIT-G using the general axioms with the corresponding SIT. The figure shows four plots for different values of n and τ of the mean

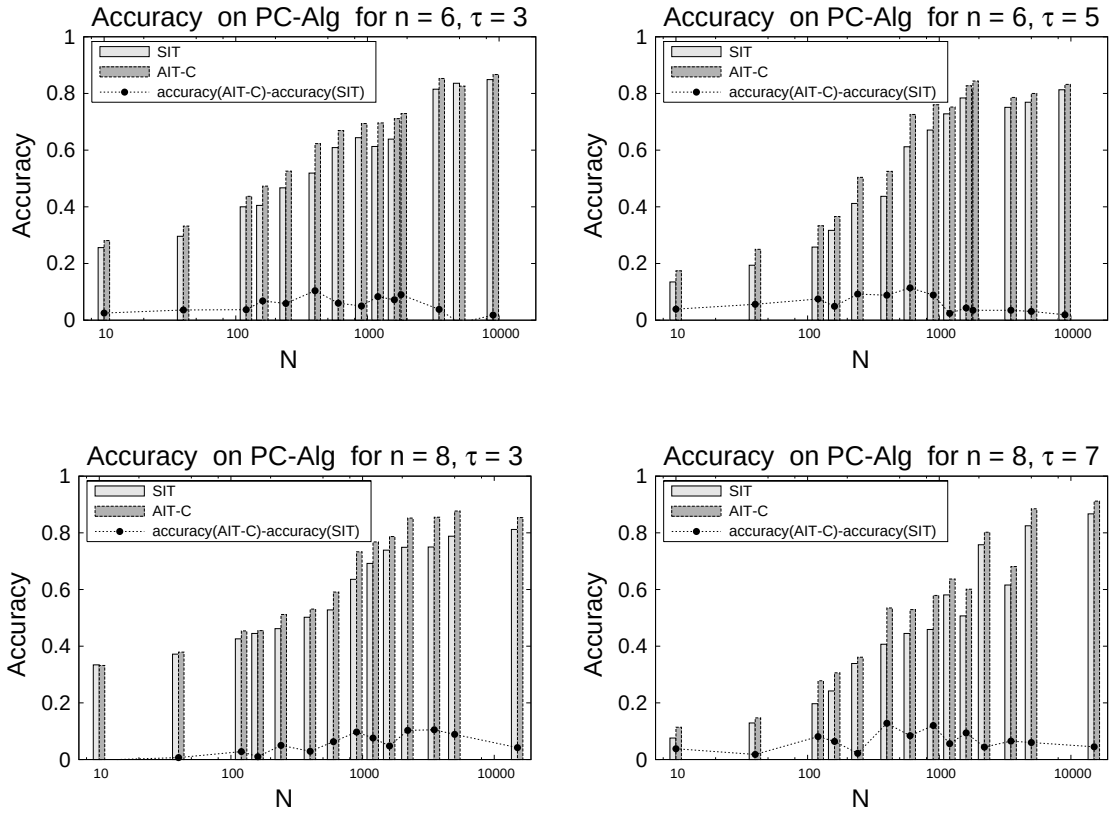

Figure 5: Comparison of the output of PC-Algorithm for SIT and AIT-C for domain size $n = 3$, maximum degree $\tau = 3, 5$ and domain size $n = 8$, maximum degree $\tau = 3, 7$. The histogram shows the absolute value of the accuracies and the line curve shows their difference i.e., a positive value correspond to an improvement in the accuracy of AIT-C over the accuracy of SIT.

values (over runs for ten different networks) of $\widehat{accuracy}^{\text{data}}_{\text{SIT}}$ and $\widehat{accuracy}^{\text{data}}_{\text{AIT}-\text{G}}$ (histograms), and the difference $(\widehat{accuracy}^{\text{data}}_{\text{AIT}-\text{G}} - \widehat{accuracy}^{\text{data}}_{\text{SIT}})$ (line graph) for different data set sizes N. A positive value of the difference corresponds to an improvement of AIT-G over SIT. We can observe modest improvements over the entire range of data set sizes in all four cases of up to 6% for $n = 6$, $\tau = 5$ and $N = 240$.

In certain situations it may be the case that the experimenter knows that the underlying distribution is causal i.e., it belongs to the class of Bayesian networks. In these situations it is appropriate to use the causal axioms of Eq. (6) instead of the general axioms of Eq. (5). Figure 3 compares the argumentative test AIT-C that uses the causal axioms vs. statistical tests. The plots follow the same format as Figure 2, with histograms plotting the mean values of $\widehat{accuracy}^{\text{data}}_{\text{SIT}}$ and $\widehat{accuracy}^{\text{data}}_{\text{AIT}-\text{C}}$ and the line graphs showing the difference $(\widehat{accuracy}^{\text{data}}_{\text{AIT}-\text{C}} - \widehat{accuracy}^{\text{data}}_{\text{SIT}})$. As in the case for the AIT using the general axioms, we can observe improvements over the entire range of data set sizes in all four cases. In this case however, the improvement is considerably larger, with sustained increases in the accuracy in the order of 5% and above, and improvement of up to 17% for $n = 8$, $\tau = 7$ and $N = 600$. We also notice in both AIT-G and AIT-C that larger improvements tend to appear in more connected domains i.e., for larger values of τ.

We also studied the effect that the improvement in the accuracy of argumentative tests has on the discovery of the structure of Bayesian networks. In the following experiments we used

the PC algorithm (Spirtes et al., 2000), an independence-based algorithm, to learn the structure. We compared the true structure of the underlying model to the resulting structure of the PC algorithm when it uses SITs as independence tests, denoted PC-SIT, and its output when it uses argumentative independence tests, denoted PC-AIT-G and PC-AIT-C when using general and causal axioms respectively. We evaluated the resulting networks by their ability to accurately represent the true independences in the domain, estimated by comparing the results (`true` or `false`) of a number of conditional tests conducted using d-separation, with the results on the output networks (PC-SIT, PC-AIT-G or PC-AIT-C). Denoting by $\mathcal{T}$ this set of 2,000 triplets, by $t \in \mathcal{T}$ a triplet, by $I_{\text{true}}(t)$ the result of a test performed on the underlying model, and by $I_{\text{PC-}\mathcal{Y}}(t)$ the result of performing a d-separation test on the output network PC-$\mathcal{Y}$ with $\mathcal{Y}$ equal to SIT, AIT-G or AIT-C, the estimated accuracy of network PC-$\mathcal{Y}$ is defined as

$$\widehat{accuracy}_{\mathcal{Y}}^{\text{PC}} = \frac{1}{|\mathcal{T}|}\left|\left\{t \in \mathcal{T} \mid I_{\text{PC-}\mathcal{Y}}(t) = I_{\text{true}}(t)\right\}\right|.$$

The comparison of the accuracy of the PC algorithm using SITs vs. using argumentative tests AIT-G or AIT-C is shown in Figures 4 and 5, respectively. The figures contain four plots each for the different values of n and τ, and have the same format as in previous figures. Once again, all four plots show improvements of the argumentation approach over the entire range of data set sizes, with improvements of up to 8% for the general axioms (for $n = 6$, $\tau = 5$ and $N = 400$), and up to 17% for the causal axioms (for $n = 8$, $\tau = 7$ and $N = 400$ and 900).

4.2 Real-world Data Experiments

While the sampled data set studies of the previous section have the advantage of a more controlled and systematic study of the performance of the algorithms, experiments on real-world data are necessary for a more realistic assessment. Real data are also more challenging because it is frequently not known whether there exists a faithful Bayesian network for the probability distribution of domain i.e., our assumptions may be violated.

We conducted experiments on a number of real-world data sets obtained from the UCI machine learning repository (D.J. Newman and Merz, 1998) and the Knowledge Discovery Data repository (Hettich and Bay, 1999). For each data set D, we conducted experiments on subsets d of D containing an increasing number of data points N. In this way we could assess the performance of the independence tests (SITs or AITs) on varying conditions of reliability, as the reliability of a test varies (typically increases) with the amount of data available. To reduce variance, each experiment was repeated for ten subsets d of equal size, obtained by permuting the data points of D randomly and using the first N as the subset d.

Because for real-world data sets the underlying model is unknown, we can only be sure the general axioms of Eq. (5) apply. Therefore in the following experiments we only report the accuracy of AIT-G, the argumentative independence test defined over the general axioms. The accuracy for this set of experiments is now defined as

$$\widehat{accuracy}_{\mathcal{Y}}^{\text{data}} = \frac{1}{|\mathcal{T}|}\left|\left\{t \in \mathcal{T} \mid I_{\text{data-}\mathcal{Y}}(t) = I_{\text{true}}(t)\right\}\right|$$

where $\mathcal{Y}$ is equal to wither SIT or AIT-G. Unfortunately, since the underlying model is not known, it is also impossible to know the true value I_{true} of any independence t. We therefore approximate it by a statistical test on the entire data set, and limit the size N of the data set subsets that we use to a third of the size of the entire data set. This corresponds to the hypothetical scenario that

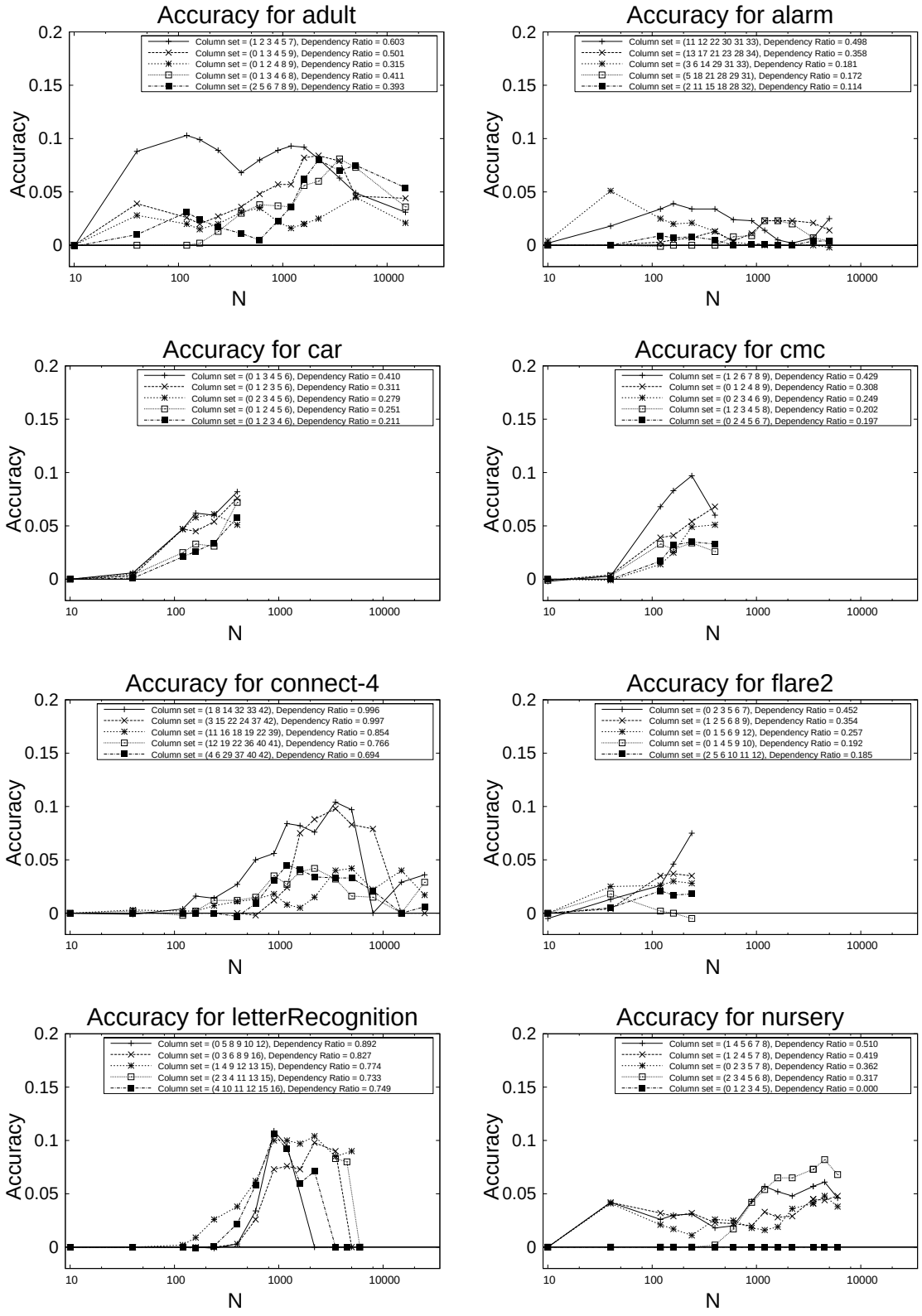

Figure 6: Accuracy improvements of AIT over SIT for a number of 6-column projections of several real-world data sets.

Table 1: Accuracies (in percentage) of SIT and AIT-G (denoted AIT in the table) for several 6-variable projections of real-world data sets. For each data set projection, the table shows the ratio of dependencies in the data set and the accuracy for number of data points $N = 40, 240, 600, 1200, 3500$. The best performance between SIT and AIT is indicated in bold. Blank table entries correspond to cases where the original data set was smaller than the value of N in that column.

Data set				N=40		N=240		N=600		N=1200		N=3500	
Name	total_N	Dep-Ratio	Variable set	SIT	AIT	SIT	AIT	SIT	AIT	SIT	AIT	SIT	AIT
alarm	20000	0.498	(11 12 22 30 31 33)	59.4	**61.2**	73.3	**76.7**	81.1	**83.5**	85.9	**87.3**	89.2	**89.9**
		0.358	(13 17 21 23 28 34)	64.7	64.7	74.0	**74.7**	77.7	**78.1**	81.5	**83.8**	91.5	**93.6**
		0.181	(3 6 14 29 31 33)	90.9	**96.0**	96.8	**98.9**	98.4	**98.7**	98.7	98.7	98.6	98.6
		0.172	(5 18 21 28 29 31)	86.4	86.4	90.2	90.2	90.4	**91.2**	92.1	**94.4**	96.2	**96.9**
		0.114	(2 11 15 18 28 32)	88.9	88.9	93.8	**94.6**	95.0	95.0	95.0	**95.1**	95.2	**95.6**
adult	32560	0.603	(1 2 3 4 5 7)	42.8	**51.6**	54.6	**63.5**	63.1	**71.1**	69.3	**78.6**	80.4	**86.7**
		0.501	(0 1 3 4 5 9)	51.4	**55.3**	56.9	**59.6**	60.9	**65.7**	66.8	**72.5**	74.5	**82.4**
		0.315	(0 1 2 4 8 9)	71.1	**73.9**	75.0	**77.0**	77.8	**81.3**	82.2	**83.8**		
		0.411	(0 1 3 4 6 8)	58.9	58.9	59.7	**61.0**	61.7	**65.5**	64.7	**68.3**	71.4	**79.5**
		0.393	(2 5 6 7 8 9)	62.1	**63.1**	65.4	**67.1**	67.2	**67.7**	69.3	**72.9**	75.3	**82.3**
nursery	12959	0.510	(1 4 5 6 7 8)	50.8	**55.0**	58.1	**61.2**	63.8	**65.8**	68.7	**74.4**	82.7	82.7
		0.419	(1 2 4 5 7 8)	60.6	**64.8**	66.6	**69.8**	70.6	**72.8**	74.0	**77.3**	84.0	84.0
		0.362	(0 2 3 5 7 8)	66.4	**70.5**	71.0	**72.1**	73.7	**76.2**	76.3	**77.9**	85.6	85.6
		0.317	(2 3 4 5 6 8)	68.3	68.3	68.3	68.3	69.3	**71.0**	72.4	**77.8**	83.5	83.5
		0.000	(0 1 2 3 4 5)	100.0	100.0	99.9	99.9	100.0	100.0	100.0	100.0	100.0	100.0
connect-4	65534	0.996	(1 8 14 32 33 42)	**0.9**	0.8	8.0	**9.4**	13.1	**18.1**	24.6	**33.0**	52.0	**62.4**
		0.997	(3 15 22 24 37 42)	0.7	0.7	0.4	0.4	**1.2**	1.0	6.3	**8.7**	31.3	**41.1**
		0.854	(11 16 18 19 22 39)	19.5	**19.8**	23.1	**23.8**	25.6	**26.9**	32.6	**33.4**	43.1	**47.1**
		0.766	(12 19 22 36 40 41)	23.9	**24.1**	25.9	**27.1**	32.7	**34.2**	39.2	**41.9**	58.0	**61.2**
		0.694	(4 6 29 37 40 42)	31.2	31.2	31.2	31.2	33.8	**34.7**	39.6	**44.1**	55.3	**58.6**
letter-rec	19999	0.892	(0 5 8 9 10 12)	10.8	10.8	**11.5**	11.4	13.8	**17.2**	21.9	**31.4**		
		0.827	(0 3 6 8 9 16)	17.3	17.3	17.3	17.3	19.2	**21.8**	26.1	**33.7**	52.1	**61.1**
		0.774	(1 4 9 12 13 15)	22.6	22.6	23.7	**26.3**	26.9	**33.1**	32.2	**42.2**	54.1	**62.6**
		0.734	(2 3 4 11 13 15)	26.6	26.6	28.6	28.6	31.6	31.6	37.8	37.8	57.6	57.6
		0.749	(4 10 11 12 15 16)	25.1	25.1	25.0	25.0	26.8	26.8	33.3	33.3	-100.0	-100.0
car	1727	0.410	(0 1 3 4 5 6)	60.1	**60.7**	69.7	**75.7**						
		0.311	(0 1 2 3 5 6)	69.8	**70.1**	77.9	**83.3**						
		0.279	(0 2 3 4 5 6)	73.3	**73.8**	80.7	**86.8**						
		0.251	(0 1 2 4 5 6)	75.2	**75.6**	80.4	**83.5**						
		0.211	(0 1 2 3 4 6)	79.1	**79.2**	83.4	**86.8**						
cmc	1472	0.429	(1 2 6 7 8 9)	58.3	**58.6**	68.9	**78.6**						
		0.308	(0 1 2 4 8 9)	70.1	**70.5**	76.6	**82.0**						
		0.249	(0 2 3 4 6 9)	**75.6**	75.5	78.0	**82.9**						
		0.202	(1 2 3 4 5 8)	81.1	**81.4**	89.0	**92.4**						
		0.197	(0 2 4 5 6 7)	79.9	79.9	84.1	**87.6**						
flare2	1065	0.452	(0 2 3 5 6 7)	62.8	**64.1**	81.5	**89.0**						
		0.354	(1 2 5 6 8 9)	67.6	**68.0**	82.9	**86.4**						
		0.257	(0 1 5 6 9 12)	79.4	**81.9**	89.2	**92.0**						
		0.192	(0 1 4 5 9 10)	83.4	**85.2**	**89.1**	88.6						
		0.185	(2 5 6 10 11 12)	82.4	**82.9**	86.6	**88.4**						

a much smaller data set is available to the researcher, allowing us to evaluate the improvement of argumentation under these more challenging situations.

As mentioned in the beginning of the experimental section, because of the exponential nature of our algorithm we has to limit the size of our domain. For real-world data sets we limited our experiments to 6 variables by selecting multiple random subsets of 6 variables from each data set D, resulting in a number of projections of D of size 6. As noted in the sampled data experiments of the previous section, the amount of improvement in accuracy is greater for more connected models, as measured by the maximum degree parameter τ used to create the underlying model. For this reason we investigated analogous situations for real-world data sets as well. As for the latter the underlying model is unknown, no connectivity parameter τ is available; instead we used as measure of dependence the ratio of the triplets that are dependent (obtained using a statistical independence test) in a collection of tests, and generated and evaluated a number of data set projections of various different ratios.

Table 1 shows the results of our comparison between argumentative tests AIT-G and statistical tests SIT for real-world data sets. The best-performing method (SIT or AIT-G, the latter shown abbreviated as AIT in the table) is shown in bold. As we can verify, the argumentative test improves accuracy for most data set sizes with the exception of very small data sets i.e., 10 data points. The same numbers are plotted in Figure 6. The figure contains 8 plots, one per data set D, each containing 5 curves for each of the variable projections of D. The plots depict the average of the difference between the accuracy of AIT-G and that of SIT, where as usual a positive value denotes an improvement of AIT-G over SIT. The figure demonstrates a clear advantage of the argumentative approach, with all data sets reaching positive improvements in accuracy of up to 10%.

5 Conclusions and Future Research

We presented a framework for addressing one of the most important problems of independence-based structure discovery algorithms, namely the problem of unreliability of statistical independence tests. Our main idea was to recognize that there exist constraints in the outcome of conditional independence tests—in the form of Pearl's axiomatic characterization of the conditional independence relation—that can be exploited to correct unreliable statistical tests. We modeled this setting as a knowledge base containing conditional independences that are potentially inconsistent, and used the preference-based argumentation framework to reason with and possibly resolve these inconsistencies. We presented in detail how to apply the argumentation framework to independence knowledge bases and how to compute the preference among the independence propositions. Our experimental results on sampled and real-world data sets show improvements in the number of correct tests (as measured by accuracy on independences) for an overwhelming majority of situations considered, with maximum improvements of up to 17% in certain cases.

As our main concern was to investigate accuracy improvements, we did not address the issue of efficiency in this paper. The development of efficient algorithms for inference using the preference-based argumentation framework in independence knowledge bases is an interesting and useful topic of future research.

Bibliography

A. Agresti. *Categorical Data Analysis*. Wiley, 2nd edition, 2002.

L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34:197–215, 2002.

W. G. Cochran. Some methods of strengthening the common χ^2 tests. *Biometrics*, 10:417–451, 1954.

C. B. D.J. Newman, S. Hettich and C. Merz. UCI repository of machine learning databases. *Irvine, CA: University of California, Department of Information and Computer Science*, 1998.

P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

P. Gärdenforst. *Belief Revision*. Cambridge Computer Tracts. Cambridge University Press, Cambridge, 1992.

P. Gärdenforst and H. Rott. Belief revision. In Gabbay, D. M., Hogger, C. J. and Robinson, J. A., editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 4. Clarendon Press, Oxford, 1995.

S. Hettich and S. D. Bay. The UCI KDD archive. *Irvine, CA: University of California, Department of Information and Computer Science*, 1999.

J. S. Ide, F. G. Cozman, and F. T. Ramos. Generating random bayesian networks with constraints on induced width. *Brazilian Symposium on Artificial Intelligence, Recife, Pernambuco, Brazil*, 2002.

M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA, 1994.

R. P. Loui. Defeat among arguments: a system of defeasible inference. *Computational Intelligence*, 2:100–106, 1987.

J. P. Martins. Belief revision. In Shapiro, S. C., editor, *Encyclopedia of Artificial Intelligence*, pages 110–116. John Wiley & Sons, New York, second edition, 1992.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA, 2nd edition, 1988.

J. L. Pollock. How to reason defeasibly. *Artificial Intelligence*, 57:1–42, 1992.

H. Prakken. *Logical Tools for Modelling Legal Argument. A Study of Defeasible Reasoning in Law.* Kluwer Law and Philosophy Library, Dordrecht, 1997.

H. Prakken and G. Vreeswijk. *Logics for Defeasible Argumentation*, volume 4 of *Handbook of Philosophical Logic*. Kluwer Academic Publishers, Dordrecht, 2 edition, 2002.

S. C. Shapiro. Belief revision and truth maintenance systems: An overview and a proposal. Technical Report CSE 98-10, Dept of Computer Science and Engineering, State University of New York at Buffalo, 1998.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Adaptive Computation and Machine Learning Series. MIT Press, 2nd edition, January 2000.